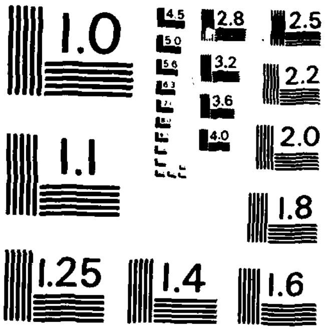
AD-A142 248   NEW APPROACHES TO ROBUST CONFIDENCE INTERVALS FOR   1/1
LOCATION: A SIMULATION STUDY(U) EDUCATIONAL TESTING
SERVICE PRINCETON NJ H I BRAUN ET AL. JUN 84
UNCLASSIFIED   ARO-18484.1-MA DAAG29-81-K-0178   F/G 12/1   NL

END



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ARO 18484.1-MA

2. GOVT ACCESSION NO.
AD-A142 248

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
NEW APPROACHES TO ROBUST CONFIDENCE INTERVALS FOR LOCATION: A SIMULATION STUDY

5. TYPE OF REPORT & PERIOD COVERED
Final/ Sept 21 1981-Sept. 30 1983

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Henry I. Braun
Daryl Pregibon, ATT Bell Laboratories
Murray Hill, N.J. 07974

8. CONTRACT OR GRANT NUMBER(s)
DAAG29-81-K-0178

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Educational Testing Service
Princeton, New Jersey 08541

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
May , 1984

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Robustness of Validity, Permutation Test, Conditional Inference
Likelihood Function

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The robustness of validity of four methods for setting confidence intervals for a location parameter when the scale is unknown are investigated. Three methods involve estimating the variance of an M-estimate of location while the fourth is a procedure suggested by Maritz, based on a permutation argument. The first three methods use either a finite sample approximation to the asymptotic variance (a well known standard) or make inferences on the basis of the shape of the putative likelihood function. The latter approach is related to the work of Sprott, as well as

20.

that of Efron and Hinkley on conditional inference. Overall, the Maritz procedure performs best though the standard does surprisingly well.

New Approaches to Robust Confidence Intervals for Location: A Simulation Study

Final Report

Henry Braun

Educational Testing Service

and

Daryl Pregibon

AT & T Bell Laboratories

June 1984

Research sponsored by a contract with U.S. Army Research Office No. DAAG29-81-K-0178, awarded to Educational Testing Service.

Approved for public release; distribution unlimited.

Acknowledgements

The authors would like to thank Bruce Kaplan for his extensive programming assistance throughout the project. In addition, Donna Lembeck's coordination and preparation of this report deserves our sincerest appreciation.

Table of Contents

Page

Abstract

1. Introduction 1

2. Review of the Literature 2

3. Procedures to Be Studied 6

4. The Data 15

5. The Simulation 15

6. Results 20

7. Concluding Remarks 22

References 25

Tables

Table 1. Coverage Probabilities for 80 and 95 Percent Nominal Confidence Intervals

Table 2. Conditional 95 Percent Coverage Probabilities

Table 3. Medial Values of Pivotal Rescaling Factor, $\bar{c}(\underline{x})$

Figure

Figure 1. Plot of Relative Likelihood and Two Approximations for a Sample from Slash (2df) Using Estimates Based on a Student's t(2df)

Abstract

The robustness of validity of four methods for setting confidence intervals for a location parameter when the scale is unknown are investigated. Three methods involve estimating the variance of an M-estimate of location while the fourth is a procedure suggested by Maritz, based on a permutation argument. The first three methods use either a finite sample approximation to the asymptotic variance (a well-known standard) or make inferences on the basis of the shape of the putative likelihood function. The latter approach is related to the work of Sprott, as well as that of Efron and Hinkley on conditional inference. Overall, the Maritz procedure performs best though the standard does surprisingly well.

Key words: Robustness of Validity, Permutation Test, Conditional Inference, Likelihood Function

1. Introduction

Most of the research in the area of robustness has focused on the problem of point estimation and comparatively little attention has been paid to the companion problems of interval estimation and testing. The present work has been directed at one of the simplest realistic testing problems: The construction of confidence intervals for a location parameter when samples of independent observations are drawn from a symmetric parent distribution of unknown location and scale. The goal is to achieve validity robustness in small samples. That is, we seek procedures for which the empirical coverage probabilities agree with the nominal values over a range of sampling distributions.

Our particular interest has centered on adapting some ideas of Sprott and his workers on "small-sample asymptotics" in the classical parametric setting (Sprott and Kalbfleisch (1969), Sprott (1973), Sprott (1980)) to the robustness problem. The essential idea is to use the shape of the likelihood function as an indication of the distribution of the maximum likelihood estimate. However, in practical robustness problems, the true form of the likelihood function is unknown and our investigation was prompted by a desire to see whether the putative likelihood function (implicit in the choice of the estimation technique) contained useful information as well.

Our numerical results suggest that while Sprott's original idea carries over quite well to the robustness setting, his particular implementation does not, and other methods are required. One such method is developed and its performance compared both with that of a standard procedure (see Gross, 1976) and a novel permutation approach of Maritz (1979). Moreover, some connections with the work of Efron and Hinkley on conditional inference are explored (Efron and Hinkley (1978), Hinkley (1978)).

2. Review of the Literature

There are a number of different strands that can be identified in the fabric of research in this area. One of the oldest begins with Gayen (1949) who investigated the distribution of Student's t-statistic when the data are not normally distributed. His corrections are functions of the skewness and kurtosis of the parent. In practice, these moments must be estimated from the data and for small samples this is quite impractical. More recently, Yuen and Murthy (1974) have studied the special case of sampling from a member of the t-family and have derived empirically compact approximations to the distribution of the usual t-statistic. The usefulness of this work has been diminished by the advances made by Hampel and his coworkers (Hampel (1973), Field and Hampel (1982)) in the area of "small-sample asymptotics."

Another approach has been based on Huber's M-estimator. The notion is to obtain a robust estimate of location and a corresponding estimate of its variability from which a t-like statistic may be constructed. Intuitively, the notion is that such a statistic should display reasonable robustness of validity. For example, Gross (1976) carried out extensive empirical investigations of the behavior of some 25 different statistics under four different sampling distributions. The most successful ones employed either the jackknife or a finite-sample version of the asymptotic standard deviation of the estimator to obtain a denominator for the test statistic. Those statistics based on location estimates derived from Hampel's redescending influence function or Tukey's bisquare performed quite well, both in terms of robustness of validity and robustness of efficiency.

An interesting proposal was put forward by Maritz (1979). He pointed out that classical permutation arguments could be used in obtaining confidence intervals for M-estimates of location. A difficulty arises when the scale parameter is unknown because some estimates of scale destroy the vaidity of the permutation argument. However other common robust scale estimates such as the median absolute deviation are permissible. Because Maritz's procedure conditions on the absolute deviations of the observations from the center of the distribution, the resulting confidence limits should be both conditionally as well as unconditionally exact. The question of conditional confidence levels arises in the work of Efron and Hinkley (1978) and Hinkley (1978) and will be treated in Section 3 below. Unfortunately, Maritz did not carry out any empirical studies.

Boos (1980) has proposed a procedure, motivated by a solution to the problem of quantile estimation, which may be thought of as a simple approximation to the more complex Maritz procedure. It seems also to be related to a suggestion made by Bickel (1976) on carrying out robust analyses by applying classical methods to appropriately transformed data. The Boos and Maritz procedures are fairly demanding computationally and the former suffers from the additional constraint that only non-decreasing ψ-functions, can be employed, thus ruling out Hampel's redescending ψ-functions. Boos carried out a small empirical study and concluded that his procedure held a small advantage in robustness of validity and efficiency over the usual studentization method.

The above review has centered on methods involving the construction of a test statistic from a particular estimator. A more ambitious plan is to formulate an optimality criterion for testing in the robustness framework and to develop test statistics meeting this criterion. The censored likelihood ratio test of Huber (1965) is an early example of a test of two simple hypotheses with specific robustness properties. Ylvisaker (1977) and later Lambert (1981) proposed different approaches. In particular, Lambert defined an influence function for a test in terms of the behavior of its P-values when the data are sampled from a model distribution modified by point contamination. Lambert examines the influence functions of a number of common test procedures.

Schrader and Hettmansperger (1980) introduced likelihood ratio-type tests, based on robust loss functions, for the general linear model. As is customary in the robustness literature, these authors advocate joint estimation of location and scale parameters but fix the latter at its estimated value for hypothesis testing for the former. Specifically, following work of Huber (1967), they show that the asymptotic distribution of the difference in maximized robust loss functions from a full model to a reduced model is proportional to $\chi^2(m)$, where m is the corresponding reduction in dimensionality. The only effect of mismatching the robust loss functions to the distribution of the data is in the constant of proportionality. Although these authors did not explicitly demonstrate how their proposal could be used for interval estimation as well as hypothesis testing, the extension is immediate.

Once an influence function has been defined, considerations similar to those proposed by Hampel (1974) in the estimation case can be brought to bear in the construction of new test statistics. A direct extension of Hampel's influence function to the testing arena can be found in Ronchetti (1979) and Rousseeuw and Ronchetti (1981). Ronchetti (1982) discusses the connection between this influence function and others that have been proposed and suggests that an appropriate optimality criterion for a test of a simple null hypothesis is to maximize the asymptotic power (within a given class of tests, appropriately standardized) subject to a fixed bound on the influence function at the null hypothesis. Ronchetti (1982) also discusses other notions such as the change-of-variance function, which is germane to estimation and a change-of-power function which is germane to testing. Optimal test statistics are derived, though their small sample properites are not investigated.

Interestingly, Rieder (1978) and Millar (1983) both propose a simple test statistic similar to that of Boos and prove certain asymptotic optimality properties.

3. Procedures to Be Studied

Our work on deriving confidence intervals for robust estimates of location stems originally from some ideas proposed by Sprott and Kalbfleisch (1965, 1969), reviewed in Sprott (1973, 1975) and extended in Sprott (1980).

Sprott and Kalbfleisch considered the usual application of maximum likelihood theory in the construction of confidence intervals in a classical parameteric setting. Suppose n independent observations are taken from a distribution F in the family $\{F_\theta\}$ indexed by a parameter θ. Under regularity conditions, the MLE of θ, $\hat{\theta}$, is asymptotically normally distributed; i.e.,

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\theta)) \qquad (3.1)$$

where $I(\theta)$ is the Fisher information. In practice, confidence statements for θ based on $\hat{\theta}$ are made on the basis of (3.1). However, when n is small, the asymptotic theory may not apply and the actual coverage probability may differ considerably from the nominal level. Following Fisher (1922), Sprott and Kalbfleisch suggest that the shape of the likelihood function for θ may give some indication of the applicability of the asymptotics. In particular, if the observed likelihood differs greatly from the normal likelihood implied by (3.1) (e.g., by being highly asymmetric) then the appropriateness of the usual interval is doubtful.

To be considered as a meaningful statement in a frequentist sense, the last sentence must be recast as follows: Suppose we identify an infinite set of samples generating likelihoods $\{L(\theta)\}$ under the model F_θ which differ trivially from each other but all very different from the corresponding functions of θ: $[2\pi I^{-1}(\theta)]^{-1/2} \exp\{-\frac{1}{2}[(\theta - \underset{\sim}{\hat{\theta}})I(\theta)]^2\}$ $\equiv L_N(\theta)$. (The quantity $\hat{\theta}_n$ changes with each sample). Then the claim is that the proportion of the resulting intervals that cover the true value of θ will differ from the nominal level; i.e., the conditional level of the procedure will be incorrect.

They illustrate this principle with an example based on sampling from the exponential distribution with density $\theta^{-1} \exp(-t/\theta)$. The likelihood function for θ in small samples tends to be asymmetric and the inferences based on the asymptotics are misleading. A solution is proposed, however.

Working for convenience in terms of the relative likelihood function $R(\theta) = L(\theta)/L(\hat{\theta})$, they suggest finding a reparameterization of θ for which the relative likelihood function is more nearly normal. The usual asymptotic theory should be applied on the transformed scale and the resulting confidence interval then transformed back to the original θ scale. In the exponential case, for example, the transformation $\lambda = \theta^{-1/3}$ works well. The relative likelihood $R(\lambda) = L(\lambda)/L(\hat{\lambda})$ tend to be quite normal looking (very little asymmetry) and the actual confidence level matches the nominal level. Thus the relevance of the usual asymptotics seem not to be a function of the sample size per se but rather of the shape of the (relative) likelihood functions generated.

Sprott (1973) expands on this approach by carrying out a formal Taylor series expansion of log R(θ) about $\hat{\theta}$:

$$\log R(\theta) = -\frac{1}{2}(\theta - \hat{\theta})^2\, I(\theta) + \frac{1}{6}(\theta - \hat{\theta})^3 \frac{\partial^3}{\partial\hat{\theta}^3} \log R(\theta) + \ldots \,. \tag{3.2}$$

The first term on the RHS of (3.2) corresponds to the relative likelihood implicit in (3.1), which we denote by $R_N(\theta)$. If the second term on the RHS of (3.2) is generally nonneglibible then the use of $R_N(\theta)$ alone as a basis for inference is suspect. On the other hand, if a transformation $\lambda = \lambda(\theta)$ can be found for which the second term is generally very small, or, ideally, identically zero, then the use of the asymptotics is better justified.

Rewriting (3.2) we have

$$\log R(\theta) = \frac{1}{2}(\theta - \hat{\theta})^2 I_\theta \{1 - \frac{1}{3}(\theta - \hat{\theta})^3 I_\theta^{-1} \frac{\partial^3}{\partial\hat{\theta}^3} \log R(\theta)\} + \ldots \,.$$

If attention is confined to the region $|\theta - \hat{\theta}| < k/I^{1/2}(\theta)$, then it seems sensible to define a measure of deviation from normality by

$$F_3(\hat{\theta}) = I^{-3/2}(\theta) \frac{\partial^3}{\partial\hat{\theta}^3} \log R(\theta) \quad . \tag{3.3}$$

Under the transformation $\lambda = \lambda(\theta)$, Sprott showed that

$$F_3(\hat{\lambda}) = F_3(\hat{\theta}) + 3I^{-1}(\theta)\,\frac{d^2\lambda}{d\theta^2}\left(\frac{d\lambda}{d\theta}\right)^{-1} . \tag{3.4}$$

If λ can be found to make $F_3(\hat{\lambda})$ zero or more nearly so than $F_3(\hat{\theta})$, then the normal approximation should work better on the λ-scale.

Sprott also employs some results of Welch and Peers (1963) to explicate the connection between normal relative likelihoods and the approximate normality of the maximum likelihood estimate. Briefly, standard asymptotics implies that

$$v(\theta) = (\hat{\theta} - \theta)\, I^{1/2}(\theta)$$

is approximately distributed as a standard normal deviate. Welch and Peers showed that

$$Z(\hat{\theta},\theta) = v(\theta) + \frac{1}{6}\,(v^2(\theta) + 2)\, F_3(\hat{\theta}) + h(\theta,\hat{\theta}) ,$$

where h is complicated function of θ and θ, is more nearly a standard normal. Sprott argues that any transformation $\lambda = \lambda(\theta)$ that reduces $|F_3|$ will make the resulting $v(\lambda)$ more nearly a linear function of $Z(\lambda,\lambda)$ and hence, improve the accuracy of the normal approximation to the distribution of $v(\lambda) = (\hat{\lambda} - \lambda)I^{1/2}(\lambda)$.

Sprott (1980) extends these ideas to the case when nuisance parameters are present. Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ be a random sample of n independent observations from a distribution F in the family $\{F_{\underline{\theta}}\}$, indexed by a vector parameter $\underline{\theta} = (\theta_1,\ldots,\theta_k)$. The density of F is denoted by f. The problem is to estimate θ_1 in the presence of the nuisance parameters $\theta_2,\ldots,\theta_k$. The relative maximum likelihood function of θ_1 is defined to be

$$R_M(\theta_1;\underline{X}) = f(\underline{X};\underline{\theta}^*)/f(\underline{X};\hat{\underline{\theta}})$$

where $f(\underline{X};\underline{\theta}) = \prod_i f(x_i,\theta)$ (the likelihood of the data as a function of θ), $\hat{\underline{\theta}}$ = MLE of $\underline{\theta}$ and $\underline{\theta}^* = (\theta_1,\theta_2^*,\ldots,\theta_k^*)$ is the restricted MLE of $\underline{\theta}$ for a given value of θ_1. In the following, notational dependence on $\underline{X}$ will be suppressed.

Let $L = \log f(\underline{X};\underline{\theta})$ and define

$$I_{\alpha\beta}(\hat{\underline{\theta}}) = \left.\frac{-\partial^2 L}{\partial\theta_\alpha \partial\theta_\beta}\right|_{\underline{\theta}=\hat{\underline{\theta}}}, \qquad I_{\alpha\beta\gamma}(\hat{\underline{\theta}}) = \left.\frac{\partial^3 L}{\partial\theta_\alpha \partial\theta_\beta \partial\theta_\gamma}\right|_{\underline{\theta}=\hat{\underline{\theta}}},$$

$$I_{\alpha\beta\gamma\delta}(\hat{\underline{\theta}}) = \left.\frac{\partial^4 L}{\partial\theta_\alpha \partial\theta_\beta \partial\theta_\gamma \partial\theta_\delta}\right|_{\underline{\theta}=\hat{\underline{\theta}}} .$$

One approach to inference about θ_1 is to focus on the so-called pivotal quantity

$$u(\theta_1) = (\hat{\theta}_1 - \theta_1)[I^{11}(\hat{\underline{\theta}})]^{-1/2} .$$

It can be shown that

$$\left.\frac{\partial^2 \log}{\partial\theta_1^2} R_M(\theta_1)\right|_{\underline{\theta}=\hat{\underline{\theta}}} = -(I^{11})^{-1} ,$$

so that the Taylor expansion of log $R_M(\theta_1)$ about $\hat{\theta}_1$ yields

$$\log R_M(\theta_1) = -\frac{1}{2} u^2(\theta_1)\{1 - (\theta_1 - \hat{\theta}_1)\left(\frac{\partial^3 \log R_M}{\partial\hat{\theta}_1^3}\right)I^{11}(\hat{\underline{\theta}})/3$$

$$- (\theta_1 - \hat{\theta}_1)^2\left(\frac{\partial^4 \log R_M}{\partial\hat{\theta}_1^4}\right)I^{11}(\hat{\underline{\theta}})/12 + \ldots\} . \qquad (3.5)$$

The quality of the quadratic approximation $-\frac{1}{2} u^2(\theta_1)$ to log $R_M(\theta_1)$ depends on the magnitude of the nonconstant terms in the curly brackets on the RHS of (3.5) If we set $\theta_1 - \hat{\theta}_1 = \pm\ 3(I^{11}(\underline{\theta}))^{1/2}$ in order to confine attention to a plausible range of values for θ and substitute into (3.5), we obtain

$$\log R_M(\theta_1) = -\frac{1}{2} u^2(\theta_1)\{1 \pm F_3(\hat{\underline{\theta}}) - (3/4)\ F_4(\hat{\theta}) + \ldots\ \}$$

where

$$F_3(\hat{\theta}) = \frac{\partial^3 \log R_M(\theta_1)}{\partial\hat{\theta}_1^3}\ [I^{11}(\hat{\underline{\theta}})]^{\ 3/2}$$

$$F_4(\hat{\theta}) = \frac{\partial^4 \log R_M(\theta_1)}{\partial \hat{\theta}_1^4} [I^{11}(\hat{\underline{\theta}})]^2.$$

As in the single parameter problem, if a transformation $\lambda = \lambda(\theta_1)$ can be found to reduce F_3 (and F_4 as well), the resulting relative maximum likelihood function will tend to be more normal in repeated samples and the normal approximation to the distribution of the pivotal quantity u more defensible.

In one of his examples, Sprott (1980) considers the case of sampling from the t-family of location and scale distributions. Because of the population symmetry, $F_3(\hat{\underline{\theta}})$ tends to be quite small but $\hat{F}_4(\underline{\theta})$ may not be negligible unless the sample size is fairly large. To deal with this problem, Sprott suggests a simple device. Instead of approximating R_M by a normal curve, a t-curve should be used to account for the fact that in small samples F_4 tends to be positive indicating less precision in the data.

Now the value of F_4 at the mode of the relative maximum likelihood function of a t-distribution on M degrees of freedom is $6(M + 1)^{-1}$. Thus the approximating t-curve to log R_M is found by solving the equation $F_4(\hat{\underline{\theta}}) = 6(M + 1)^{-1}$. Denote the solution by $\hat{M} = M(\hat{\theta})$. Following the logic enunciated above, we would then suppose that $u(\theta_1)$ is distributed approximately as $t_{\hat{M}}$ and set the appropriate confidence limits for θ_1.

Sprott's work is based on the conjecture that R_M contains useful information about the behavior of u and he simply provides a convenient way of extracting the information in R_M by a simple approximation using a tabled distribution. A computationally burdensome approach would involve a more detailed look at R_M along the lines suggested by Fraser (1976).

Although the above discussion has been carried out in the classical setting, it is perfectly feasible to apply in the robustness setting. In the language of M-estimation, suppose we choose a ψ-function corresponding to a model distribution F. Data are obtained from an unknown distribution, denoted G. A (pseudo) relative maximum likelihood function is constructed based on the assumed model F and the observed data from G. The shape of this function is approximated by an appropriate t-curve and the latter is used to set confidence limits for the parameter of interest. For example, we may choose F to be the location-scale family for t_2 which corresponds to choosing a reasonable redescending ψ-function for estimation, while we may sample from a member of the slash family (Rogers and Tukey, 1977). Empirical studies must determine whether the pseudo-relative maximum likelihood function does indeed carry useful information.

In some respects, Sprott's work is closely related to that of Efron and Hinkley (1978) and Hinkley (1978) on conditional likelihood inference. They argue that in general the observed information rather than the expected information (Fisher information) is a better guide to the variability of the maximum likelihood estimate, conditional on an appropriate ancillary statistic. In the case of a single location parameter, Efron and Hinkley, building on Fisher's work (1934) on likelihood inference, show that the conditional distribution, $f_\theta(\hat{\theta}|a)$, of the MLE $\hat{\theta}$ of θ given the ancillary statistic of order statistics spacings is proportional to the likelihood function of θ. Thus, a normal shaped likelihood does indeed imply a conditional normal distribution of $\hat{\theta}$. Moreover, the variance of this conditional normal distribution is $i^{-1}(\hat{\theta})$, the reciprocal of the observed information.

Hinkley extends this result to the location-scale case. The joint conditional distribution of the MLEs given the appropriate ancillary is again proportional to the likelihood function. The conditional distribution of the pivot $(\hat{\theta}_1 - \theta_1)/\hat{\theta}_2$ is asymptotically normal with mean 0 and variance i^{11}. In an example based on sampling from the Cauchy distribution, Hinkley demonstrates the superiority of the observed information to the usual Fisher information as a measure of the variability of the pivot, though the final recommendation is to set confidence limits through a direct approximation of the likelihood function.

Thus Sprott, Efron and Hinkley all use the shape of the observed likelihood function to provide some indication of the appropriate measure of variability to be attached to a point estimate of location. They focus on how the shape of the likelihood function may invalidate the application of the usual unconditional asymptotics. Efron and Hinkley suggest conditioning as a remedy while Sprott argues that transformations are a better way of dealing with the problem. However, when an appropriate transformation cannot be found or applied, Sprott does consider alternative ways of approximating the likelihood function.

4. The Data

In the simulations conducted for this study, data from the t-family and slash family were generated. Standard pseudo-random number generators of uniform and unit normal deviates were employed. Denoting them by u and n respectively let $s = n/u^{1/\nu}$. Then s is said to follow the slash distribution with ν degrees of freedom provided n and u have been generated independently. Variates from the t-family were generated using the "ratio-of-uniforms" method described in Kinderman and Monahan (1977).

5. The Simulation

The simulation study investigates the properties of four procedures: (1) AST--A standard procedure based on the asymptotic studentization of an M estimate; (2) Tt--the procedure proposed here based on an extension of Sprott's work and employing a t-curve approximation to the observed pseudo-likelihood based on matching fourth derivatives; (3) Tn--as (2) above except that a normal approximation to the observed pseudo-likelihood is used; (4) M--the Maritz procedure.

All four procedures generate confidence intervals based on the use of the same ψ-function corresponding to the choice of model. The first three require calculation of the appropriate (robust) estimates of location and scale. The E-M algorithm (Dempster, Laird, and Rubin, 1977) was used. In order to implement Tt, all derivatives of the logarithm of the pseudo-relative maximum likelihood function up to fourth order must be computed in order that the quantity $F_4(\hat{\theta})$ can be obtained easily (see Sprott, 1980, p. 516-517). More detailed descriptions of the procedures follow below.

(1) AST - A $100(1 - \alpha)$ confidence interval is given by

$$(\hat{\theta}_1 - t_{n-1}(\alpha/2)\hat{S}n^{-1/2}, \hat{\theta}_1 + t_{n-1}(\alpha/2)\hat{S}n^{-1/2}) ,$$

where

$$\hat{S} = [(n - 1)^{-1} \sum_i \psi^2(\frac{x_i - \hat{\theta}_1}{\hat{\theta}_2})]^{1/2}\hat{\theta}_2/n^{-1} \sum_i \psi'(\frac{x_i - \hat{\theta}_1}{\hat{\theta}_2})$$

and

$t_\nu(\alpha) = 100(1 - \alpha)$ percent point of Student's-t on ν degrees of freedom.

<u>Remark</u>: An intuitive interpretation of AST is obtained by considering the pivotal $(\hat{\theta}_1 - \theta_1)/\sqrt{v}\ (\hat{\theta})$, where

$$v(\hat{\theta}) = \hat{\theta}_2/\Sigma\ \psi'(\frac{x_i - \hat{\theta}_1}{\hat{\theta}_2}) .$$

When the putative likelihood function defined by ψ matches the underlying distribution, this pivotal is approximately distributed as a standard normal variate. When these functions do not match, $v(\hat{\theta})$ misestimates the variance of $\hat{\theta}_1$ and the pivotal must be rescaled for approximate standard normality to be obtained. The correction used in AST is given by

$$c_\psi(\underline{x}) = \Sigma\ \psi'\left(\frac{x_i - \hat{\theta}_1}{\hat{\theta}_2}\right)/\Sigma\ \psi^2\left(\frac{x_i - \hat{\theta}_1}{\hat{\theta}_2}\right)$$

so that $v(\hat{\theta})/c_\psi(\underline{x})$ is the appropriate variance estimate of θ_1. This interpretation will prove useful when we contrast the methods below.

(2) Tt - A $100(1 - \alpha)$ confidence interval is given by

$$(\hat{\theta}_1 - t_M(\alpha/2)[i^{11}(\hat{\underline{\theta}})]^{1/2},\ \hat{\theta}_1 + t_M(\alpha/2)[i^{11}(\hat{\underline{\theta}})]^{1/2})\ ,$$

where M is the solution to $F_4(\hat{\underline{\theta}}) = 6(M + 1)^{-1}$.

(3) Tn - A $100(1 - \alpha)$ confidence interval is given by

$$(\hat{\theta}_1 - z(\alpha/2)[i^{11}(\hat{\underline{\theta}})/\bar{c}]^{1/2},\ \hat{\theta}_1 + z(\alpha/2)[i^{11}(\hat{\underline{\theta}})/\bar{c}]^{1/2})\ ,$$

where

$z(\alpha/2) = 100(1 - \alpha)$ percent point of the standard normal distribution.

$\bar{c} = \bar{c}(\underline{x})$ = a moderating factor for the usual asymptotic variance that depends on the data observed.

Remark: Both Tt and Tn stem from the notion that the usual conditional asymptotic statement, namely that $(\hat{\theta}_1 - \theta_1)/[i^{11}(\hat{\underline{\theta}})]^{1/2}$ is distributed as a standard normal variate, is not valid in small samples. Tt uses instead a t-approximation to the observed pseudo-likelihood while Tn adjusts the asymptotic variance $[i^{11}(\hat{\underline{\theta}})]^{1/2}$, again by recourse to the observed pseudo-likelihood. In other words, Tn behaves as if the likelihood is indeed normal but with a spread that may differ substantially from that suggested by the asymptotics.

In this simulation we have used a fairly crude method to determine the adjustment factor $\bar{c}(\underline{x})$. The relative maximum likelihood function $R_M(\cdot)$ is evaluated at three pairs of points symmetrically placed about $\hat{\theta}_1$. For k = 1,2 and 3, let

$$w_k = \frac{1}{2}\,[R_M(\hat{\theta}_1 + k(i^{11}(\underline{\hat{\theta}}))^{1/2}) + R_M(\hat{\theta}_1 - k(i^{11}(\underline{\hat{\theta}}))^{1/2})] \ .$$

Then set $c_k = -2k^{-2} \log w_k$ and $\bar{c}(\underline{x}) = \frac{1}{3}\,(c_1 + c_2 + c_3)$. Thus $\bar{c}(\underline{x})$ represents a compromise among three estimates of the required scaling factor, obtained by looking at different points on the shoulders of $R_M(\cdot)$. If R_M is exactly normal then $\bar{c}(\underline{x}) = 1$.

(4) M - A 100(1 - α) confidence interval is given by (θ_L, θ_U) where θ_L and θ_U are solutions to the equations

$$\sum_i \operatorname{sgn}(x_i - t)\psi(|X_i - t|/s) = \pm\, z(\alpha/2)[\sum_i \psi^2(|x_i - t|/s)]^{1/2}$$

where $s = \text{med}\{|x_i - t|\}$.

<u>Remark</u>: The above prescription actually represents a convenient normal approximation to the full permutation distribution derived by Maritz (1979). He noted that the usual permutation argument applied to means or

nonparameteric statistics like the Wilcoxon signed rank statistic could could also be applied to M-estimates. If we define

$$M_\psi(\underline{x},t) = \Sigma\ \mathrm{sgn}(x_i - t)\psi(|x_i - t|)\ ,$$

then the M-estimate of location for the data $\underline{x}$ is the solution of the equation $M_\psi(\underline{x},t) = 0$. To obtain a $(1 - 2r/2^n)$ two-sided confidence interval (t_1,t_2), the values t_1 and t_2 must be determined by finding the r^{th} smallest and the r^{th} largest values of t solving the equations

$$\sum_i \mathrm{sgn}(x_i' - t)\psi(|x_i' - t|) = 0$$

where the summation is over $i = 1,2,\ldots,s$ and $s = 1,2,\ldots,n$. That is we consider all possible solutions to the basic equation when the data are allowed to vary over all subsets of the original data. The desired values t_1 and t_2 are the r^{th} smallest and r^{th} largest of these solutions.

In practice this calculation is somewhat demanding so that a normal approximation is recommended. Secondly, an estimate of scale is often needed and one that is a function of the absolute deviations $|x_1 - t|$, $|x_2 - t|,\ldots,|x_n - t|$ may be employed without disturbing the permutation argument.

Remark: Boos' (1980) procedure is essentially equivalent to Maritz's except that s is replaced by some fixed estimate of scale (not depending on t) and a t distribution on (n - 1) degrees of freedom, rather than the normal distribution, is employed as the reference.

6. Results

The results of the major simulations are presented in Table 1. While Tt performs adequately when the data are generated from the t-family, it breaks down quite badly when slash-family data are used. On the other hand, Tn performs quite well throughout though it is somewhat inferior to AST for slash data (but superior to AST for t_2 data). The Maritz procedure performs about as well overall as AST and somewhat better in fact for t_2 data.

Insert Table 1 about here

It seems clear that the device of approximating R_M by a t-curve based on matching fourth derivatives at the mode is too sensitive to the shape of the underlying parent distribution. Figure 1 displays a typical R_M and its approximations by t-curves and a normal curve. The t-curve approximation is quite poor while the normal curve one is excellent. Similar experiments were run for samples of size 10 and 40 but as the comparisons are qualitively the same as for size 20, they are not presented here.

Insert Figure 1 about here

The conditional coverage probabilities of the procedures are also of some interest, especially in light of the Efron-Hinkley proposals. Table 2 presents results for four combinations of data and model which which are illustrative of the results obtained for the full set of of combinations depicted in Table 1. For each combination, the coverage probabilities for the three procedures are based on the same set of 1000 samples which have been divided into thirds based on the value of $i^{11}(\hat{\underline{\theta}})$. The cut points are provided. There is an obvious pattern of increasing coverage probability with increasing $i^{11}(\hat{\underline{\theta}})$. Of course, ideally there should be no trend with $i^{11}(\hat{\underline{\theta}})$. Particularly in the case of the Tn procedure, it appears as if the low observed values of $i^{11}(\hat{\underline{\theta}})$ are "too" low while high ones are "too" high.

Employing jackknifed values of i^{11} in the construction of Tn confidence intervals immediately suggests itself as a possible remedy. However a small experiment based on jackknifing log i^{11} and then transforming back did not give promising results. While the Maritz procedure performs better than the others as one would expect from its theoretical properties, its conditional coverage probabilities do follow the same trend as those of the others.

Insert Table 2 about here

Recall now that Hinkley (1978) found that $i^{11}(\hat{\underline{\theta}})$ seriously underestimates the variance of $\hat{\theta}_1$ when sampling from the Cauchy distribution ($n = 20$). Our findings support this. In Table 3, we display the median values of $\bar{c}(\underline{x})$ for each combination of model and data, each again based on a run of 1000 samples. The quantities in parentheses are the ratios of the interquartile range to the median for each batch of 1000 values of $\bar{c}(\underline{x})$. That $\bar{c}(\underline{x})$ does not vary much across samples within a given sampling situation explains why it does not also vary across the sampling situations investigated here. In any case, the values of $\bar{c}(\underline{x})$ are substantially less than unity indicating that the use of $i^{11}(\hat{\underline{\theta}})$ alone would produce intervals much too liberal.

Insert Table 3 about here

7. Concluding Remarks

Our results agree with those of Sprott in the sense that the applicability of asymptotic normality depends most on the shape of the observed likelihood function than on sample size. Our results generalize his since our experiements demonstrate that pseudo-likelihoods, corresponding to robust M-estimates, also share this property.

Specifically, our results indicate that the small sample distribution of the pivotal

$$\frac{\theta_1 - \theta_1}{\sqrt{\underline{i}^{11}}\ (\theta)} \tag{7.1}$$

is approximately normal with mean zero, but with nonunit variance. Thus, the observed information is not a good variance estimate for $\hat{\theta}_1$ in small samples with unknown scale. This holds true whether or 'not the estimating function matches the distribution of the data.

The observed (pseudo-) likelihood function can be used to rescale the pivotal to obtain an honest small sample variance estimate. We chose to do this directly by approximating the rescaling factor by comparing the observed relative maximized likelihood function to a family of normal likelihoods. This proved quite successful *albeit* somewhat *ad hoc*. Alternative direct approximation methods might prove slightly better.

We have not been successful in determining an algebraic expression for the rescaling factor to be applied to the pivotal (7.1). Sprott's suggestion of matching the 4th derivative of an approximating t-family did not perform well. Our current conjecture is that the correct variance estimate of $\hat{\theta}_1$ when θ_2 is unknown, is the observed value of $[\underline{i}^{-1}(\hat{\theta})\underline{\Lambda}(\hat{\theta})\underline{i}^{-1}(\hat{\theta})]_{11}$ where $\underline{\Lambda}(\hat{\theta}) = \Sigma\underline{s}_i(\hat{\theta})\underline{s}_i^T(\hat{\theta})$ and $\underline{s}_i(\hat{\theta})$ is the contribution to the score vector (location and scale) of the i^{th} observation. This follows directly from Huber's (1967) general results concerning the asymptotic distribution of robust M-estimates. Our conjecture is that this expression is valid both conditionally and unconditionally, and that it should be used routinely to set confidence intervals for location parameters, even in situations where the estimator matches the distribution of the data! This variance estimate tends to correct for both small sample sizes and the mismatch of estimating function and data distribution.

Our research has identified some open problems in the areas of conditional inference and robust testing. We were aware of possible connections among these areas prior to our research reported here, and we are more strongly convinced that the results of Hinkley (1978) can be extended to situations where the putative likelihood function does not match the distribution of the data. Theoretical results along these lines would be very useful in practice since they not only provide computationally cheap alternatives to the Maritz procedure, but are easily extended to the general linear model.

A question mark also remains as to why the AST method performs so well. In particular, why joint estimation of location and scale parameters, followed by testing methods assuming known scale (fixed at its estimated value) provides such good coverage probabilities. We argue that this is a general phenomenon not well understood in the statistical community. In particular, in certain applications of Bayesian methods, fixing nuisance parameters at their estimated values rather than providing priors for them often yield extremely accurate inferences. Clearly much needs to be done to understand the mechanism behind these empirical findings.

References

Bickel, P. J. (1976) Another look at robustness: A review of reviews and some new developments. Scandinavian Journal of Statistics, 3, 145-168.

Boos, D. D. (1980) A new method for constructing approximate confidence intervals from M-estimates. Journal of the American Statistical Association, 75, 142-145.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65, 457-482.

Field, C. A. and Hampel, F. R. (1982) Small-sample asymptotic distributions of M-estimators of location. Biometrika, 69, 29-46.

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, Series A, 222, 308-358.

Fisher, R. A. (1934) Two new properties of mathematical likelihood. Proceedings of the Royal Society, London, Series A, 144, 285-307.

Fraser, D. A. S. (1976) Necessary analysis and adaptive inference. Journal of the American Statistical Association, 71, 99-113.

Gayen, A. K. (1949) The distribution of Student's t in random samples of any size drawn from non-normal universes. Biometrika, 36, 353-369.

Gross, A. M. (1976) Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association, 71, 409-416.

Hampel, F. R. (1973) Some small-sample asymptotics. Proceedings of the Prague Symposium on Asymptotic Statistics, Prague, 109-126.

Hampel, F. R. (1974) The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69, 383-393.

Hinkley, D. V. (1978) Likelihood inference about location and scale parameters. Biometrika, 65, 253-261.

Huber, P. J. (1965) A robust version of the probability ratio test. Annals of Mathematical Statistics, 36, 1753-1758.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Berkeley, Calif.: University of California Press.

Kinderman, A. J. and Monahan, J. F. (1977) Computer generation of random variables using the ratio of uniform deviates. ACM Transactions on Mathematical Software, 3, 257-260.

Lambert, D. (1981) Influence functions for testing. Journal of the American Statistical Association, 76, 649-657.

Maritz, J. S. (1979) A note on exact robust confidence intervals for location. Biometrika, 66, 163-166.

Miller, P. W. (1983) Robust tests of statistical hypotheses. Research Report. Berkeley, Calif.: Department of Statistics, University of California.

Rieder, H. (1978) A robust asymptotic testing model. Annals of Statistics, 6, 1080-1094.

Rogers, W. H. and Tukey, J. W. (1972) Understanding some long-tailed symmetrical distributions. Statistics Neerlandica, 26, 211-256.

Ronchetti, E. (1979) Robustheitseignschaften von tests. Diploma Thesis, ETH Zurich.

Ronchetti, E. (1982) Robust testing in linear models: The infinitesimal approach. Dissertation submitted to Mathematics Department, Swiss Federal Institute of Technology, Zurich.

Rousseeuw, P. J. and Ronchetti, E. (1981) Influence curves for general statistics. Journal of Computational and Applied Mathematics, 7, 161-166.

Schrader, R. M. and Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. Biometrika, 67, 93-101.

Sprott, D. A. (1973) Normal likelihoods and their relation to large sample theory of estimation. Biometrika, 60, 457-465.

Sprott, D. A. (1975) Application of maximum likelihood methods to infinite samples. Sankhyā, 37, 259-270.

Sprott, D. A. (1980) Maximum likelihood in small samples: Estimation in the presence of nuisance parameters. Biometrika, 67, 515-523.

Sprott, D. A. and Kalbfleisch, J. G. (1965) Use of the likelihood function in inference. Psychological Bulletin, 64, 15-22.

Sprott, D. A. and Kalbfleisch, J. G. (1969) Examples of likelihoods and comparison with point estimates and large sample approximations. Journal of the American Statistical Association, 64, 468-484.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. Journal of the Royal Statistical Society, Series B, 25, 318-329.

Ylvisaker, D. (1977) Test resistance. Journal of the American Statistical Association, 72, 551-556.

Yuen, K. K. and Murthy, V. K. (1974) Percentage points of the distribution of the t-statistic when the parent is Student's t. Technometrics, 16, 495-497.

Table 1

Coverage Probabilities for 80 and 95 Percent Nominal Confidence Intervals

$n = 20$

Data	Model	AST	T_t	T_n	M	AST	T_t	T_n	M
t_2	t_2	.87	.81	.82	.82	.98	.96	.96	.96
	t_5	.84	.84	.81	.77	.97	.97	.96	.95
	t_{10}	.78	.82	.84	.80	.95	.97	.96	.95
t_5	t_2	.81	.77	.79	.84	.95	.93	.94	.96
	t_5	.82	.80	.79	.80	.95	.94	.94	.95
	t_{10}	.79	.79	.80	.78	.95	.95	.93	.94
t_{10}	t_2	.80	.77	.78	.84	.95	.93	.94	.97
	t_5	.81	.80	.79	.80	.95	.95	.95	.96
	t_{10}	.80	.79	.78	.78	.95	.94	.94	.95
s_2	t_2	.82	.63	.75	.79	.96	.83	.92	.95
	t_5	.83	.76	.82	.81	.96	.92	.95	.95
	t_{10}	.80	.78	.81	.78	.95	.94	.95	.94
s_5	t_2	.76	.59	.71	.78	.92	.78	.89	.95
	t_5	.80	.71	.78	.81	.95	.90	.93	.95
	t_{10}	.80	.74	.79	.78	.95	.92	.94	.94
s_{10}	t_2	.75	.57	.71	.78	.93	.79	.89	.95
	t_5	.79	.70	.77	.80	.94	.89	.93	.94
	t_{10}	.79	.74	.79	.78	.95	.92	.93	.94

Note: Results for AST and T_t based on the same 5000 samples; results for the T_n and M based on the same 1000 samples.

Table 2

Conditional 95 Percent Coverage Probabilities

						Cut values of
Data	Model	$i^{11}(\hat{\underline{\theta}})$	AST	T_n	M	$i^{11}(\hat{\underline{\theta}})$
t_2	t_2	L	.96	.93	.94	
						.041
		M	.99	.96	.96	
						.059
		H	.99	.98	.98	
t_2	t_{10}	L	.92	.92	.92	
						.09
		M	.96	.96	.95	
						.14
		H	.97	.99	.97	
s_2	t_2	L	.95	.85	.91	
						.059
		M	.98	.95	.98	
						.087
		H	.98	.95	.97	
s_2	t_{10}	L	.92	.91	.92	
						.15
		M	.96	.97	.95	
						.18
			.96	.98	.96	

Table 3

Medial Values of Pivotal Rescaling Factor, $\bar{c}(\underline{x})$

n = 20

	t_2	t_5	t_{10}	s_2	s_5	s_{10}
t_2	.57	.54	.53	.55	.53	.52
	(.09)	(.08)	(.08)	(.095)	(.095)	(.09)
t_5	.71	.70	.69	.71	.69	.69
	(.04)	(.025)	(.02)	(.04)	(.02)	(.02)
t_{10}	.79	.78	.78	.78	.78	.78
	(.02)	(.006)	(.004)	(.01)	(.005)	(.004)

n = 10

	t_2	t_5	t_{10}	s_2	s_5	s_{10}
t_2	.54	.52	.51	.53	.51	.51
	(.11)	(.10)	(.09)	(.13)	(.10)	(.10)
t_5	.66	.66	.65	.66	.66	.65
	(.035)	(.02)	(.02)	(.04)	(.02)	(.02)
t_{10}	.73	.73	.73	.73	.73	.73
	(.004)	(.001)	(.0017)	(.003)	(.002)	(.0017)

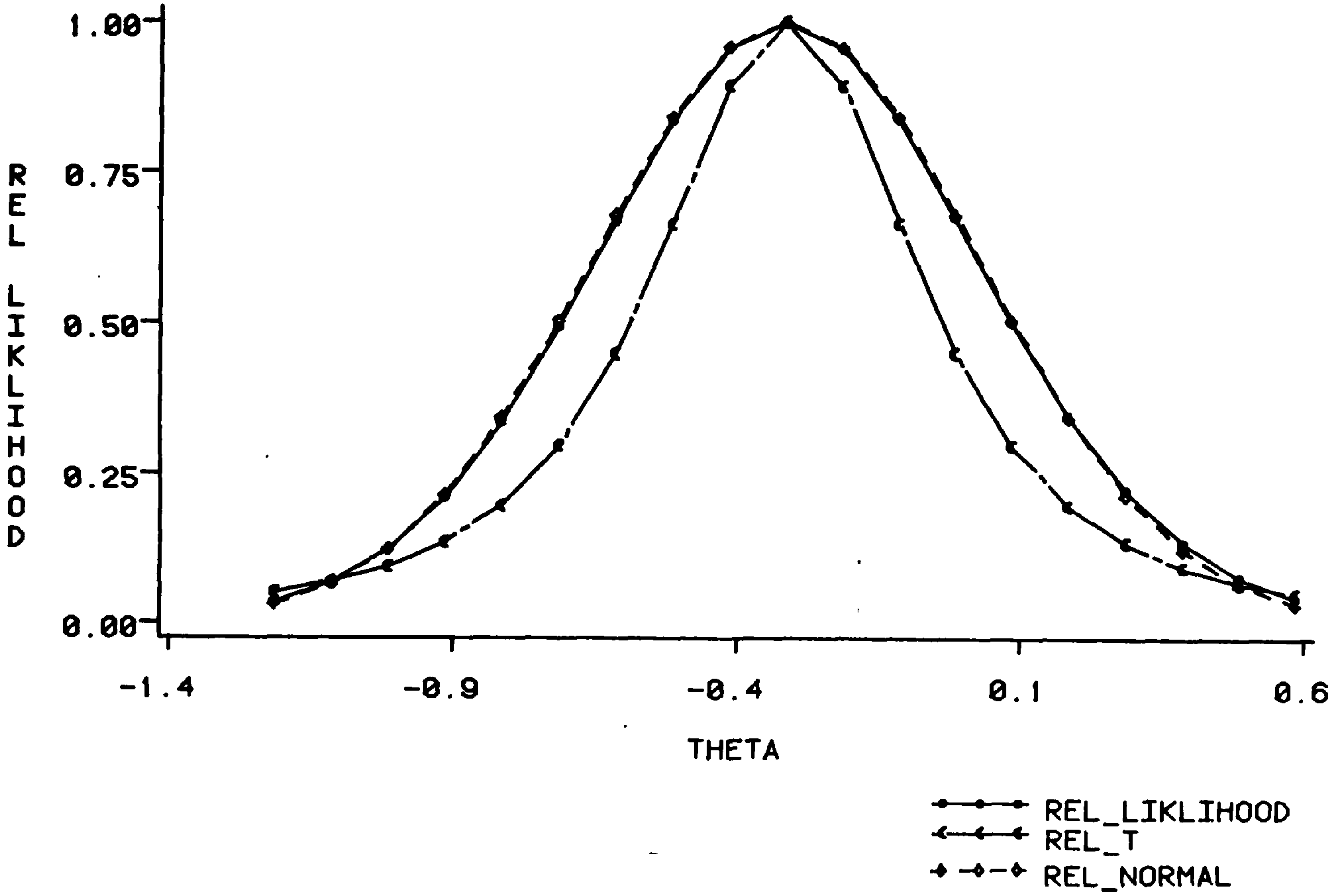


Figure 1: Plot of relative likelihood and two approximations for a sample from Slash (2df) using estimates based on a student's t (2df)